The State of Data Curation at NeurIPS: An Assessment of Dataset Development Practices in the Datasets and Benchmarks Track

Eshta Bhardwaj, Harshit Gujral, Siyi Wu, Ciara Zogheib, Tegan Maharaj, Christoph Becker
University of Toronto, Canada

Check out the project here

Key takeaway: The creation of the D&B track shows that **dataset quality is the foundation of continued progress in ML** applications. There is no better database of knowledge than data curation to aid in this venture.

Our evaluation framework provides a practical lens on how NeurIPS can spearhead the requirement for **rigorous data curation in ML**.

NEURAL INFORMATION PROCESSING SYSTEMS

Just Sustainability Design Lab

UNIVERSITY OF TORONTO
FACULTY OF INFORMATION
Digital Curation Institute

Introduction

- NeurIPS has responded to the rising urgency and recognized impact of data research through the introduction of the **D&B track**
- This track aims to address the issue of datasets being used **outside their original scope**

Background

- **Data curation** involves "maintaining and adding value to digital research data for current and future use"
- Field of data curation has **established methods and discourse** on how to maintain large amounts of data and manage ethical concerns

Motivation

- ML research has turned towards the **improvement of data to improve model results** and fundamental understanding
- Current **gap in recognition and uptake** of data curation concepts in the ML community

Our Goal

- Document and improve the standard of **dataset development in NeurIPS** so that future benchmarks and datasets can be effectively found, easily accessed, ethically used, consistently evaluated, and appropriately reused

Research Questions and Methods

What constitutes a well curated dataset?

- Developed an **evaluation framework** made up rubric and toolkit
- **Rubric** evaluates dataset contents and dataset design decisions
- **Toolkit** provides application guidance for the rubric

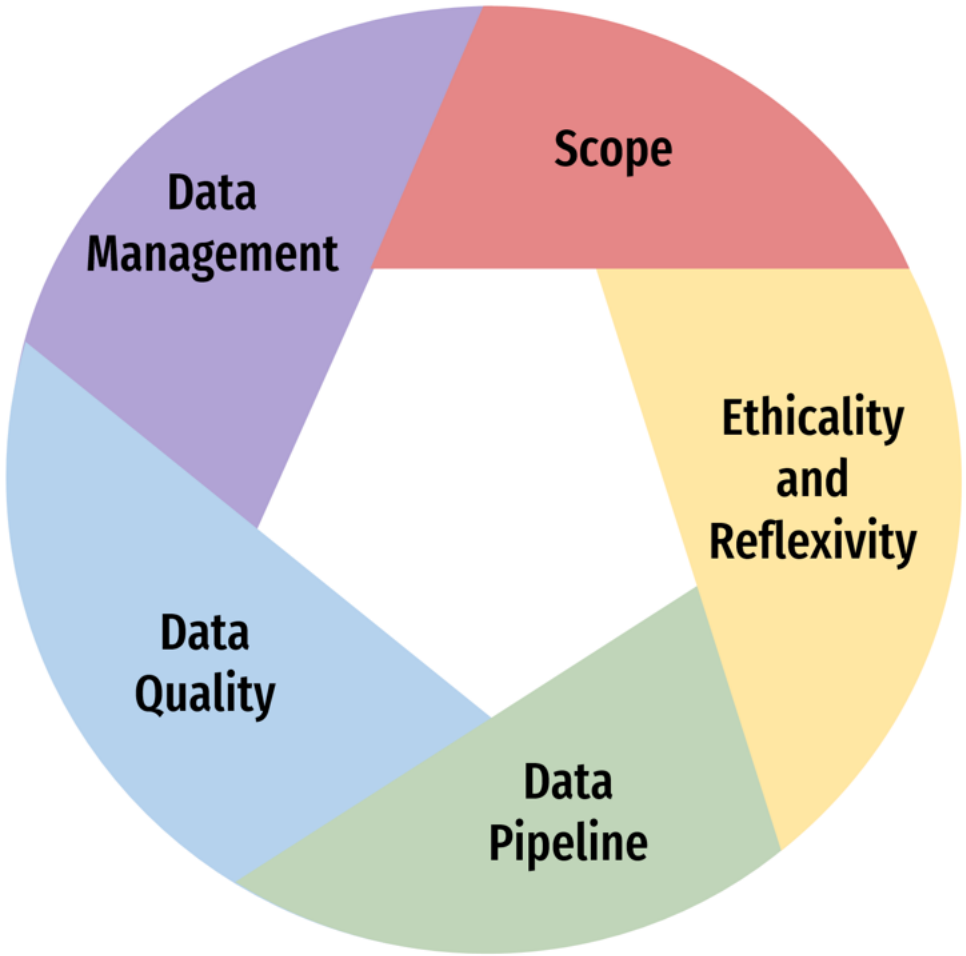
How feasible is the adoption of data curation principles to assess ML datasets?

- Applied framework to **evaluate** NeurIPS datasets
- Examined the consistency in application by measuring **inter-rater reliability** (IRR)
- Improved IRR through **iterative** rounds of evaluation and framework development

What is the state of data curation at NeurIPS?

- Assessed datasets to **evaluate current practices of data curation** in ML dataset development
- Analyzed areas in which **improvement** was needed

Evaluation Framework



Context, purpose, motivation
Requirements

Ethicality
Domain knowledge & data practices
Context awareness
Environmental footprint

Data collection
Data processing
Data annotation

Suitability
Representativeness
Authenticity
Reliability
Structured documentation

Findability
Accessibility
Interoperability
Reusability

Context awareness

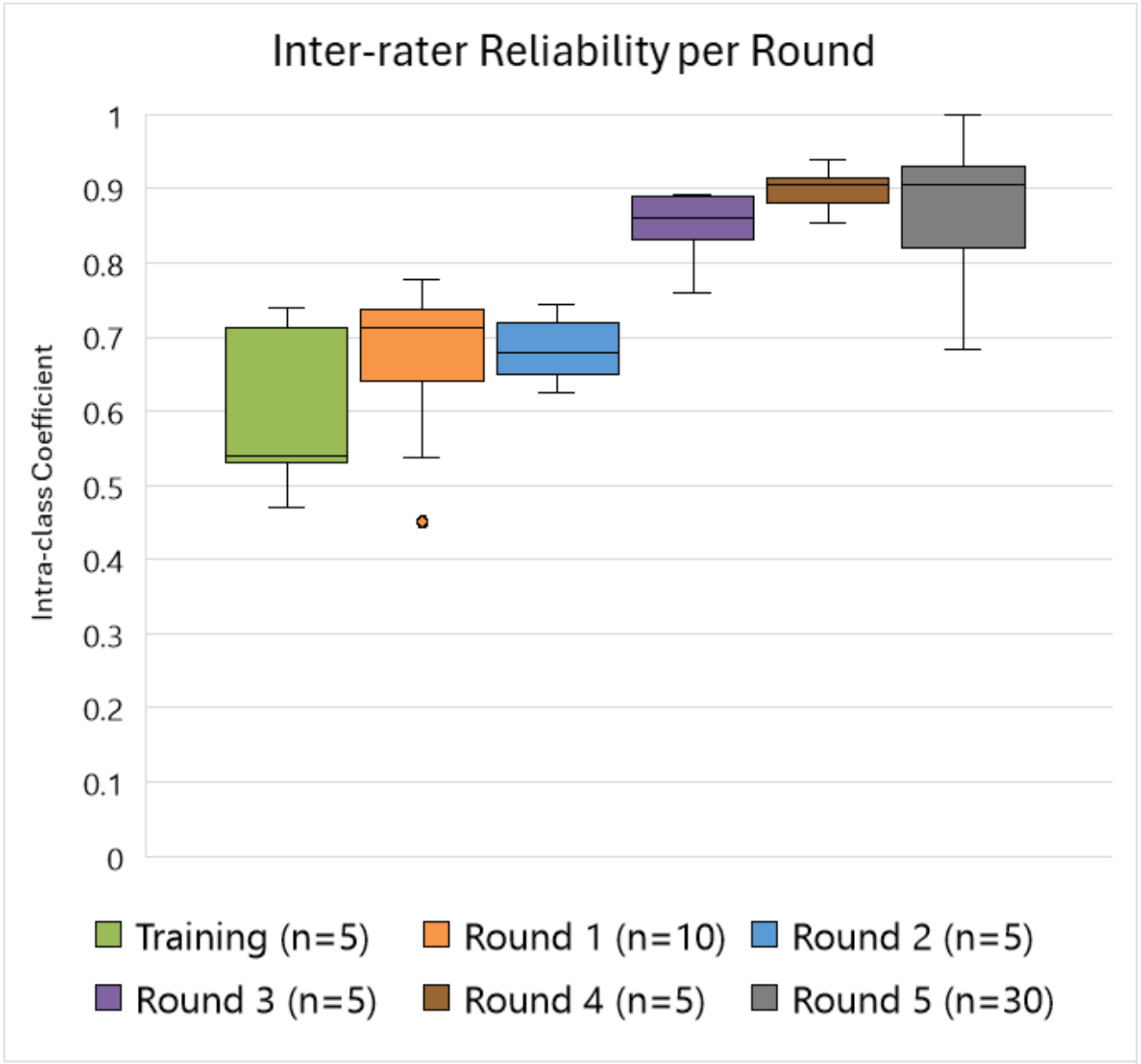
Context awareness demonstrates an understanding of the subjective, non-neutral nature, and situatedness of data.

Criteria to meet minimum standard

Documentation includes a positionality statement.

Criteria to meet standard of excellence

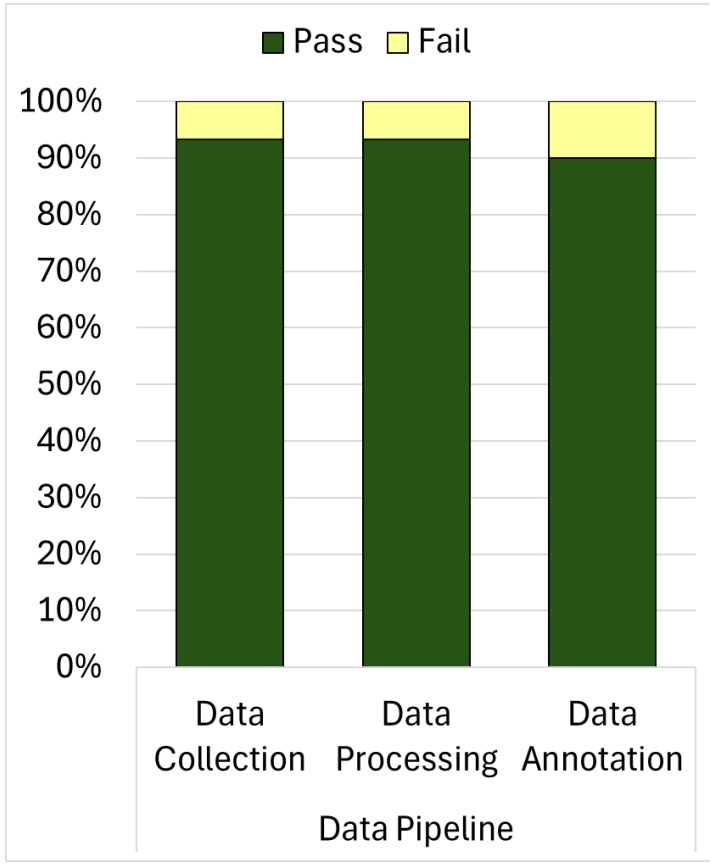
Documentation adopts a reflexive approach to dataset development. For example, documentation discusses how field epistemologies impact assumptions, methods, or framings.



Finding 1

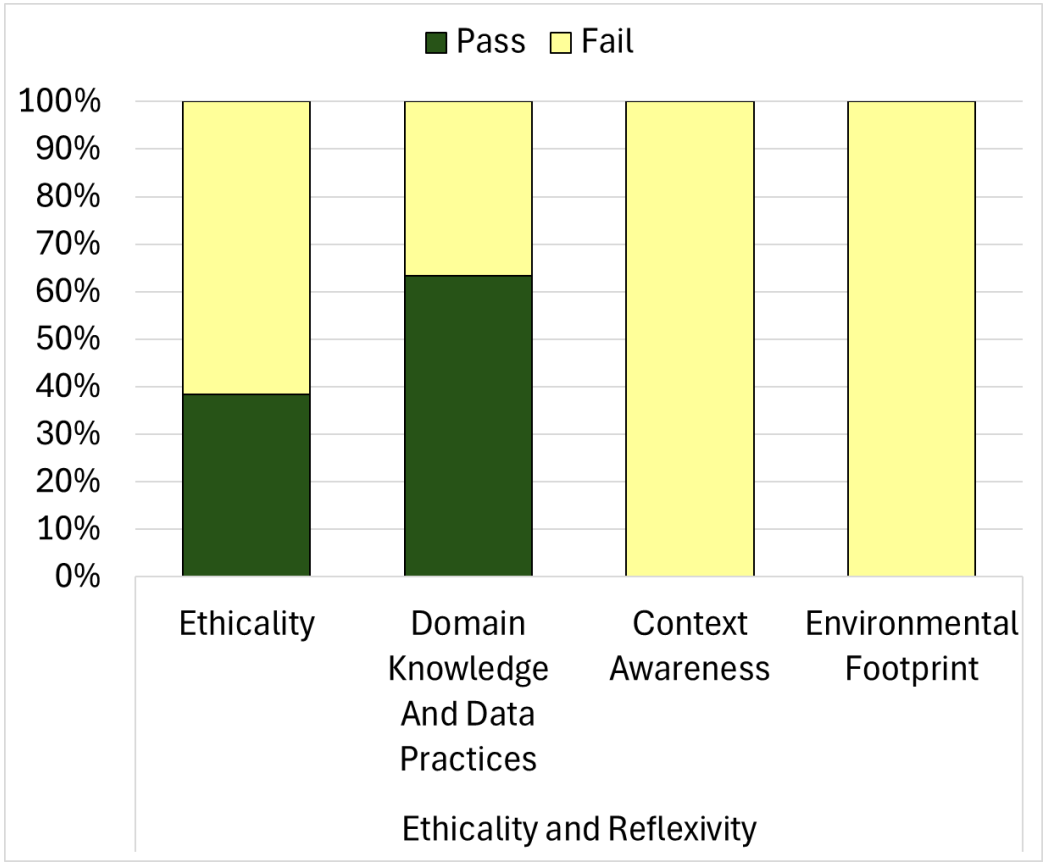
Inter-rater reliability (IRR) suggests the evaluations are consistent and reliable

Current Practices of Data Curation



Finding 2

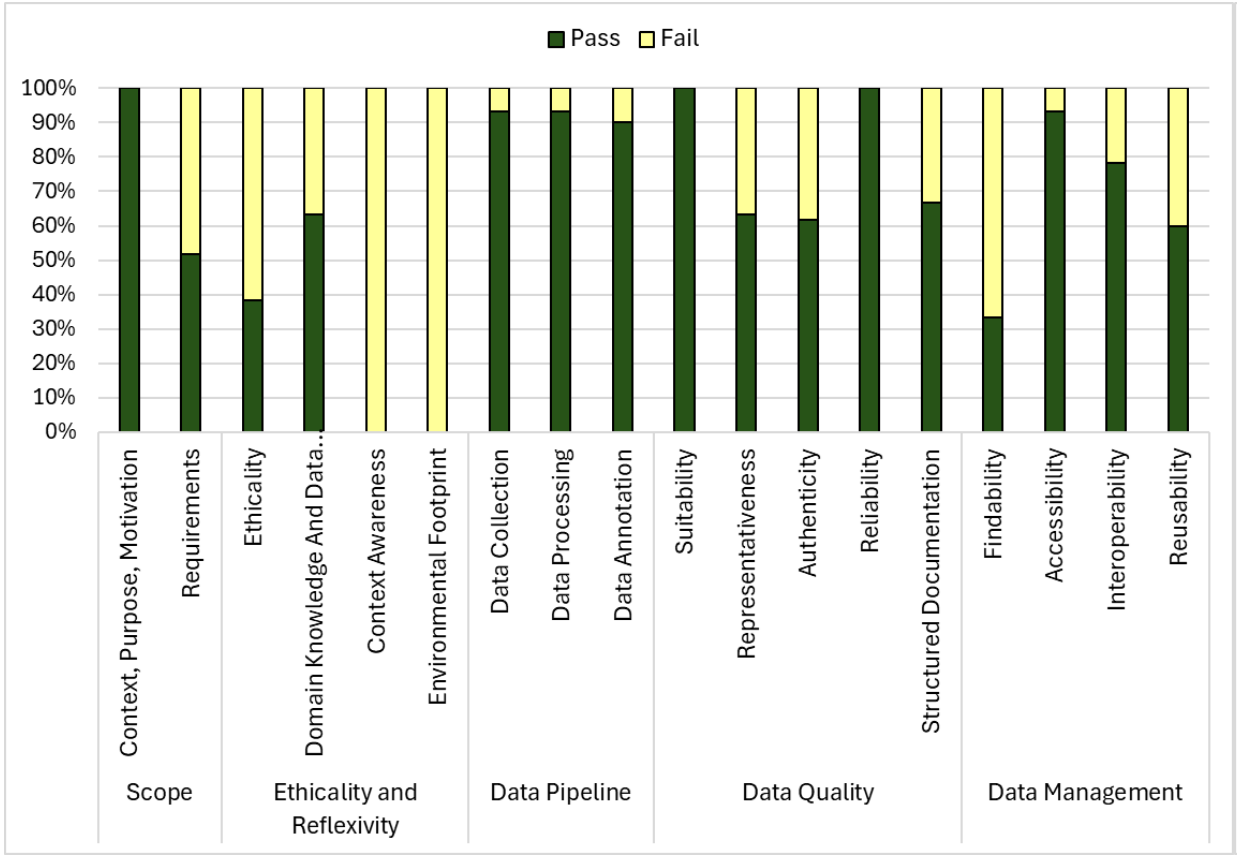
NeurIPS prioritizes model-work adjacent documentation

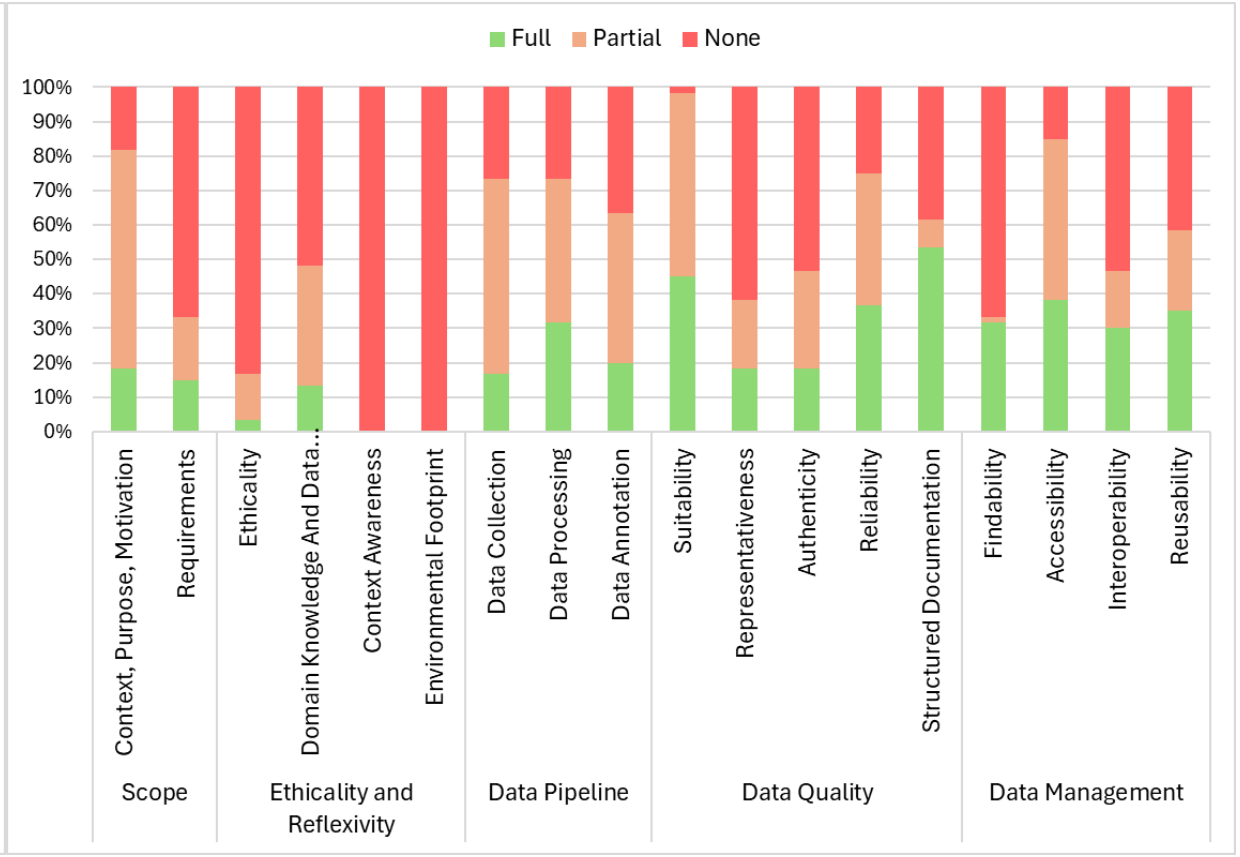


Finding 3

Documentation is rarely context aware and typically does not quantify environmental footprint

Current Practices of Data Curation



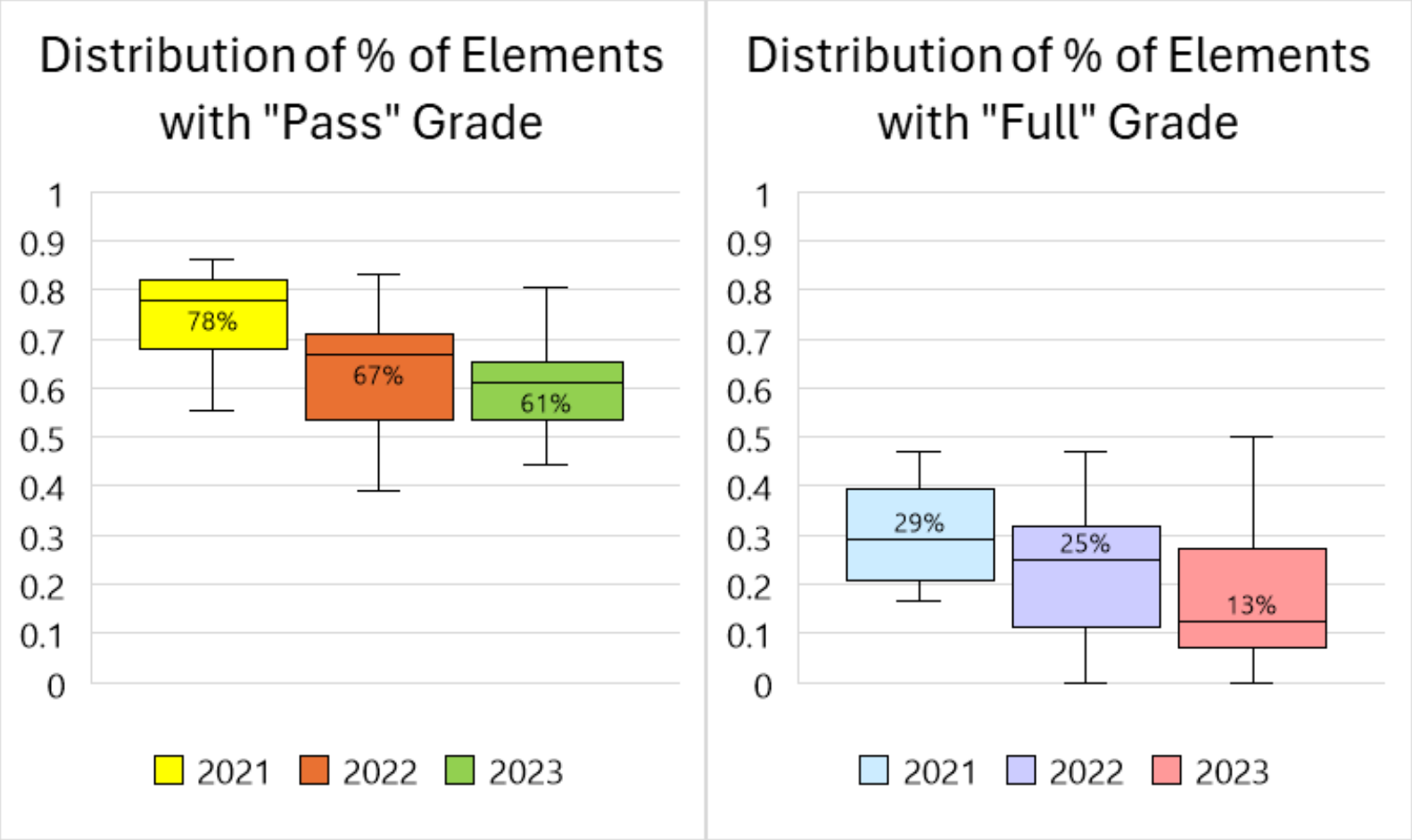


Finding 4

Documentation quality varies widely

Finding 5

Documentation often remains incomplete



Finding 6

Findings suggest no improvements occurred over time

Strategies to Improve Data Curation in ML

REQUIREMENTS

- Create **purpose statements**
- Document **initial formulation** of the problem vs. the dataset creation scheme

ETHICALITY

- Consider **proportionality principle**

CONTEXT AWARENESS

- Include **positionality statements** to increase reflexivity

ENVIRONMENTAL FOOTPRINT

- **Quantify** the environmental footprint of datasets

FINDABILITY

- Assign **persistent identifiers** to metadata to avoid link rot

REUSABILITY

- Provide dataset **provenance**

Check out the project here

Key takeaway: The creation of the D&B track shows that **dataset quality is the foundation of continued progress in ML** applications. There is no better database of knowledge than data curation to aid in this venture.

Our evaluation framework provides a practical lens on how NeurIPS can spearhead the requirement for **rigorous data curation in ML**.

Get involved by giving feedback, reach out:

Eshta Bhardwaj

Harshit Gujral

Ciara Zogheib

Siyi Wu

Tegan Maharaj

Christoph Becker